Manual and Automatic Evaluation of Machine Translation Between European Languages

Manual and Automatic Evaluation of Machine Translation between European Languages

Philipp Koehn
School of Informatics
University of Edinburgh
pkoehn@inf.ed.ac.uk

Christof Monz
Department of Computer Science
Queen Mary, University of London
christof@dcs.qmul.ac.uk

Abstract

We evaluated machine translation performance for six European language pairs that participated in a shared task: translating French, German, Spanish texts to English and back. Evaluation was done automatically using the BLEU score and manually on *fluency* and *adequacy*.

For the 2006 NAACL/HLT Workshop on Machine Translation, we organized a shared task to evaluate machine translation performance. 14 teams from 11 institutions participated, ranging from commercial companies, industrial research labs to individual graduate students.

The motivation for such a competition is to establish baseline performance numbers for defined training scenarios and test sets. We assembled various forms of data and resources: a baseline MT system, language models, prepared training and test sets, resulting in actual machine translation output from several state-of-the-art systems and manual evaluations. All this is available at the workshop website[1].

The shared task is a follow-up to the one we organized in the previous year, at a similar venue (Koehn and Monz, 2005). As then, we concentrated on the translation of European languages and the use of the Europarl corpus for training. Again, most systems that participated could be categorized as statistical phrase-based systems. While there is now a number of competitions — DARPA/NIST (Li, 2005), IWSLT (Eck and Hori, 2005), TC-Star — this one focuses on text translation between various European languages.

This year's shared task changed in some aspects from last year's:

- We carried out a manual evaluation in addition to the automatic scoring. Manual evaluation was done by the participants. This revealed interesting clues about the properties of automatic and manual scoring.
- We evaluated translation *from* English, in addition to *into* English. English was again paired with German, French, and Spanish. We dropped, however, one of the languages, Finnish, partly to keep the number of tracks manageable, partly because we assumed that it would be hard to find enough Finnish speakers for the manual evaluation.
- We included an out-of-domain test set. This allows us to compare machine translation performance in-domain and out-of-domain.

1 Evaluation Framework

The evaluation framework for the shared task is similar to the one used in last year's shared task. Training and testing is based on the Europarl corpus. Figure 1 provides some statistics about this corpus.

1.1 Baseline system

To lower the barrier of entrance to the competition, we provided a complete baseline MT system, along with data resources. To summarize, we provided:

- sentence-aligned, tokenized training corpus
- a development and development test set
- trained language models for each language
- the phrase-based MT decoder Pharaoh
- a training script to build models for Pharaoh

The performance of the baseline system is similar to the best submissions in last year's shared task. We are currently working on a complete open source implementation of a training and decoding system, which should become available over the summer.

[1] http://www.statmt.org/wmt06/

Proceedings of the Workshop on Statistical Machine Translation, pages 102–121,
New York City, June 2006.

Training corpus

	Spanish ↔ English	French ↔ English	German ↔ English
Sentences	730,740	688,031	751,088
Foreign words	15,676,710	15,323,737	15,256,793
English words	15,222,105	13,808,104	16,052,269
Distinct foreign words	102,886	80,349	195,291
Distinct English words	64,123	61,627	65,889

Language model data

	English	Spanish	French	German
Sentence	1,003,349	1,070,305	1,066,974	1,078,141
Words	27,493,499	29,129,720	31,604,879	26,562,167

In-domain test set

	English	Spanish	French	German
Sentences	2,000			
Words	59,307	61,824	66,783	55,533
Unseen words	141	206	164	387
Ratio of unseen words	0.23%	0.40%	0.24%	0.70%
Distinct words	6,031	7,719	7,230	8,812
Distinct unseen words	139	203	163	385

Out-of-domain test set

	English	Spanish	French	German
Sentences	1,064			
Words	25,919	29,826	31,937	26,818
Unseen words	464	368	839	913
Ratio of unseen words	1.79%	1.23%	2.62%	3.40%
Distinct words	5,166	5,689	5,728	6,594
Distinct unseen words	340	267	375	637

Figure 1: Properties of the training and test sets used in the shared task. The training data is the Europarl corpus, from which also the in-domain test set is taken. There is twice as much language modelling data, since training data for the machine translation system is filtered against sentences of length larger than 40 words. Out-of-domain test data is from the Project Syndicate web site, a compendium of political commentary.

ID	Participant
cmu	Carnegie Mellon University, USA (Zollmann and Venugopal, 2006)
lcc	Language Computer Corporation, USA (Olteanu et al., 2006b)
ms	Microsoft, USA (Menezes et al., 2006)
nrc	National Research Council, Canada (Johnson et al., 2006)
ntt	Nippon Telegraph and Telephone, Japan (Watanabe et al., 2006)
rali	RALI, University of Montreal, Canada (Patry et al., 2006)
systran	Systran, France
uedin-birch	University of Edinburgh, UK — Alexandra Birch (Birch et al., 2006)
uedin-phi	University of Edinburgh, UK — Philipp Koehn (Birch et al., 2006)
upc-jg	University of Catalonia, Spain — Jesús Giménez (Giménez and Màrquez, 2006)
upc-jmc	University of Catalonia, Spain — Josep Maria Crego (Crego et al., 2006)
upc-mr	University of Catalonia, Spain — Marta Ruiz Costa-jussà (Costa-jussà et al., 2006)
upv	University of Valencia, Spain (Sánchez and Benedí, 2006)
utd	University of Texas at Dallas, USA (Olteanu et al., 2006a)

Figure 2: Participants in the shared task. Not all groups participated in all translation directions.

1.2 Test Data

The test data was again drawn from a segment of the Europarl corpus from the fourth quarter of 2000, which is excluded from the training data. Participants were also provided with two sets of 2,000 sentences of parallel text to be used for system development and tuning.

In addition to the Europarl test set, we also collected 29 editorials from the Project Syndicate website[2], which are published in all the four languages of the shared task. We aligned the texts at a sentence level across all four languages, resulting in 1064 sentence per language. For statistics on this test set, refer to Figure 1.

The out-of-domain test set differs from the Europarl data in various ways. The text type are editorials instead of speech transcripts. The domain is general politics, economics and science. However, it is also mostly political content (even if not focused on the internal workings of the European Union) and opinion.

1.3 Participants

We received submissions from 14 groups from 11 institutions, as listed in Figure 2. Most of these groups follow a phrase-based statistical approach to machine translation. Microsoft's approach uses dependency trees, others use hierarchical phrase models. Systran submitted their commercial rule-based system that was not tuned to the Europarl corpus.

About half of the participants of last year's shared task participated again. The other half was replaced by other participants, so we ended up with roughly the same number. Compared to last year's shared task, the participants represent more long-term research efforts. This may be the sign of a maturing research environment.

While building a machine translation system is a serious undertaking, in future we hope to attract more newcomers to the field by keeping the barrier of entry as low as possible.

For more on the participating systems, please refer to the respective system description in the proceedings of the workshop.

2 Automatic Evaluation

For the automatic evaluation, we used BLEU, since it is the most established metric in the field. The BLEU metric, as all currently proposed automatic metrics, is occasionally suspected to be biased towards statistical systems, especially the phrase-based systems currently in use. It rewards matches of n-gram sequences, but measures only at most indirectly overall grammatical coherence.

The BLEU score has been shown to correlate well with human judgement, when statistical ma-

[2] http://www.project-syndicate.com/

chine translation systems are compared (Doddington, 2002; Przybocki, 2004; Li, 2005). However, a recent study (Callison-Burch et al., 2006), pointed out that this correlation may not always be strong. They demonstrated this with the comparison of statistical systems against (a) manually post-edited MT output, and (b) a rule-based commercial system.

The development of automatic scoring methods is an open field of research. It was our hope that this competition, which included the manual and automatic evaluation of statistical systems and one rule-based commercial system, will give further insight into the relation between automatic and manual evaluation. At the very least, we are creating a data resource (the manual annotations) that may the basis of future research in evaluation metrics.

2.1 Computing BLEU Scores

We computed BLEU scores for each submission with a single reference translation. For each sentence, we counted how many n-grams in the system output also occurred in the reference translation. By taking the ratio of matching n-grams to the total number of n-grams in the system output, we obtain the precision p_n for each n-gram order n. These values for n-gram precision are combined into a BLEU score:

$$\text{BLEU} = \text{BP} \cdot \exp(\sum_{n=1}^{4} \log p_n) \quad (1)$$

$$\text{BP} = \min(1, e^{1-r/c}) \quad (2)$$

The formula for the BLEU metric also includes a brevity penalty for too short output, which is based on the total number of words in the system output c and in the reference r.

BLEU is sensitive to tokenization. Because of this, we retokenized and lowercased submitted output with our own tokenizer, which was also used to prepare the training and test data.

2.2 Statistical Significance

Confidence Interval: Since BLEU scores are not computed on the sentence level, traditional methods to compute statistical significance and confidence intervals do not apply. Hence, we use the bootstrap resampling method described by Koehn (2004).

Following this method, we repeatedly — say, 1000 times — sample sets of sentences from the output of each system, measure their BLEU score, and use these 1000 BLEU scores as basis for estimating a confidence interval. When dropping the top and bottom 2.5% the remaining BLEU scores define the range of the confidence interval.

Pairwise comparison: We can use the same method to assess the statistical significance of one system outperforming another. If two systems' scores are close, this may simply be a random effect in the test data. To check for this, we do pairwise bootstrap resampling: Again, we repeatedly sample sets of sentences, this time from both systems, and compare their BLEU scores on these sets. If one system is better in 95% of the sample sets, we conclude that its higher BLEU score is statistically significantly better.

The bootstrap method has been critized by Riezler and Maxwell (2005) and Collins et al. (2005), as being too optimistic in deciding for statistical significant difference between systems. We are therefore applying a different method, which has been used at the 2005 DARPA/NIST evaluation.

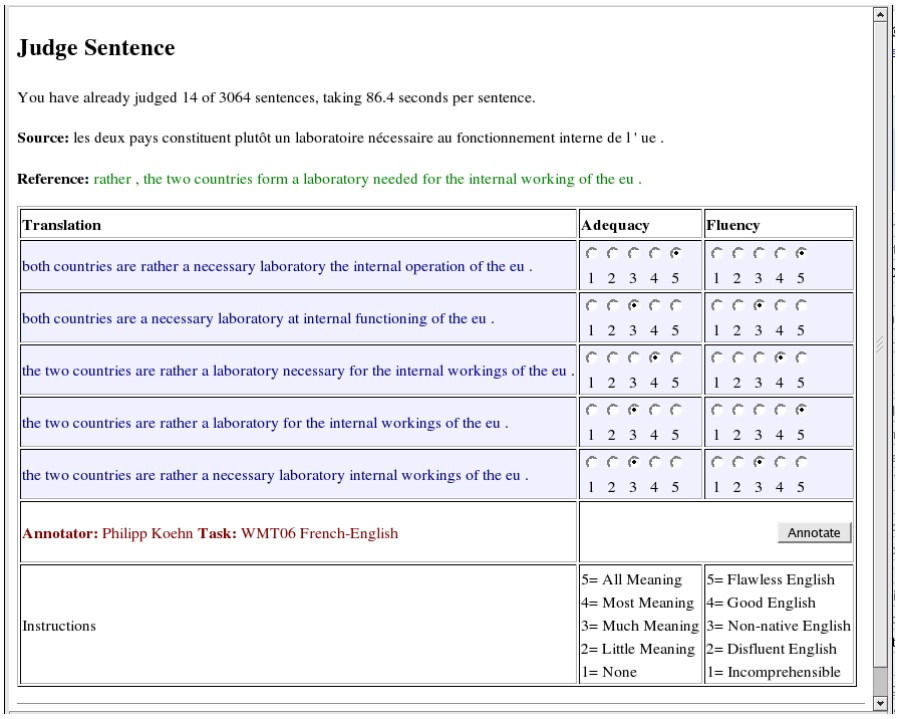
We divide up each test set into blocks of 20 sentences (100 blocks for the in-domain test set, 53 blocks for the out-of-domain test set), check for each block, if one system has a higher BLEU score than the other, and then use the sign test.

The sign test checks, how likely a sample of better and worse BLEU scores would have been generated by two systems of equal performance.

Let say, if we find one system doing better on 20 of the blocks, and worse on 80 of the blocks, is it significantly worse? We check, how likely only up to $k = 20$ better scores out of $n = 100$ would have been generated by two equal systems, using the binomial distribution:

$$\begin{aligned} p(0..k;n,p) &= \sum_{i=0}^{k} \binom{i}{n} p^i p^{n-i} \\ &= 0.5^n \sum_{i=0}^{k} \binom{i}{n} \end{aligned} \quad (3)$$

If $p(0..k;n,p) < 0.05$, or $p(0..k;n,p) > 0.95$ then we have a statistically significant difference between the systems.



Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Figure 3: Annotation tool for manual judgement of *adequacy* and *fluency* of the system output. Translations from 5 randomly selected systems for a randomly selected sentence is presented. No additional information beyond the instructions on this page are given to the judges. The tool tracks and reports annotation speed.

3 Manual Evaluation

While automatic measures are an invaluable tool for the day-to-day development of machine translation systems, they are only a imperfect substitute for human assessment of translation quality, or as the acronym BLEU puts it, a bilingual evaluation *understudy*.

Many human evaluation metrics have been proposed. Also, the argument has been made that machine translation performance should be evaluated via task-based evaluation metrics, i.e. how much it assists performing a useful task, such as supporting human translators or aiding the analysis of texts.

The main disadvantage of manual evaluation is that it is time-consuming and thus too expensive to do frequently. In this shared task, we were also confronted with this problem, and since we had no funding for paying human judgements, we asked participants in the evaluation to share the burden. Participants and other volunteers contributed about 180 hours of labor in the manual evaluation.

3.1 Collecting Human Judgements

We asked participants to each judge 200–300 sentences in terms of fluency and adequacy, the most commonly used manual evaluation metrics. We settled on contrastive evaluations of 5 system outputs for a single test sentence. See Figure 3 for a screenshot of the evaluation tool.

Presenting the output of several system allows the human judge to make more informed judgements, contrasting the quality of the different systems. The judgements tend to be done more in form of a ranking of the different systems. We assumed that such a contrastive assessment would be beneficial for an evaluation that essentially pits different systems against each other.

While we had up to 11 submissions for a translation direction, we did decide against presenting all 11 system outputs to the human judge. Our initial experimentation with the evaluation tool showed that this is often too overwhelming.

Making the ten judgements (2 types for 5 systems) takes on average 2 minutes. Typically, judges

initially spent about 3 minutes per sentence, but then accelerate with experience. Judges where excluded from assessing the quality of MT systems that were submitted by their institution. Sentences and systems were randomly selected and randomly shuffled for presentation.

We collected around 300–400 judgements per judgement type (adequacy or fluency), per system, per language pair. This is less than the 694 judgements 2004 DARPA/NIST evaluation, or the 532 judgements in the 2005 DARPA/NIST evaluation. This decreases the statistical significance of our results compared to those studies. The number of judgements is additionally fragmented by our breakup of sentences into in-domain and out-of-domain.

3.2 Normalizing the judgements

The human judges were presented with the following definition of *adequacy* and *fluency*, but no additional instructions:

	Adequacy	**Fluency**
5	All Meaning	Flawless English
4	Most Meaning	Good English
3	Much Meaning	Non-native English
2	Little Meaning	Disfluent English
1	None	Incomprehensible

Judges varied in the average score they handed out. The average fluency judgement per judge ranged from 2.33 to 3.67, the average adequacy judgement ranged from 2.56 to 4.13. Since different judges judged different systems (recall that judges were excluded to judge system output from their own institution), we normalized the scores.

The **normalized judgement per judge** is the raw judgement plus (3 minus average raw judgement for this judge). In words, the judgements are normalized, so that the average *normalized judgement per judge* is 3.

Another way to view the judgements is that they are less quality judgements of machine translation systems per se, but rankings of machine translation systems. In fact, it is very difficult to maintain consistent standards, on what (say) an adequacy judgement of 3 means even for a specific language pair.

The way judgements are collected, human judges tend to use the scores to rank systems against each other. If one system is perfect, another has slight flaws and the third more flaws, a judge is inclined to hand out judgements of 5, 4, and 3. On the other hand, when all systems produce muddled output, but one is better, and one is worse, but not completely wrong, a judge is inclined to hand out judgements of 4, 3, and 2. The judgement of 4 in the first case will go to a vastly better system output than in the second case.

We therefore also normalized judgements on a per-sentence basis. The **normalized judgement per sentence** is the raw judgement plus (0 minus average raw judgement for this judge on this sentence).

Systems that generally do better than others will receive a positive average *normalized judgement per sentence*. Systems that generally do worse than others will receive a negative one.

One may argue with these efforts on normalization, and ultimately their value should be assessed by assessing their impact on inter-annotator agreement. Given the limited number of judgements we received, we did not try to evaluate this.

3.3 Statistical Significance

Confidence Interval: To estimate confidence intervals for the average mean scores for the systems, we use standard significance testing.

Given a set of n sentences, we can compute the sample mean $\bar{x}$ and sample variance s^2 of the individual sentence judgements x_i:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (4)$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (5)$$

The extend of the confidence interval $[\bar{x}-d, \bar{x}+d]$ can be computed by

$$d = 1.96 \cdot \frac{s}{\sqrt{n}} \quad (6)$$

Pairwise Comparison: As for the automatic evaluation metric, we want to be able to rank different systems against each other, for which we need assessments of statistical significance on the differences between a pair of systems.

Unfortunately, we have much less data to work with than with the automatic scores. The way we

Basis	Diff.	Ratio
Sign test on BLEU	331	75%
Bootstrap on BLEU	348	78%
Sign test on Fluency	224	50%
Sign test on Adequacy	225	51%

Figure 4: Number and ratio of statistically significant distinction between system performance. Automatic scores are computed on a larger tested than manual scores (3064 sentences vs. 300–400 sentences).

collected manual judgements, we do not necessarily have the same sentence judged for both systems (judges evaluate 5 systems out of the 8–10 participating systems).

Still, for about good number of sentences, we do have this direct comparison, which allows us to apply the sign test, as described in Section 2.2.

4 Results and Analysis

The results of the manual and automatic evaluation of the participating system translations is detailed in the figures at the end of this paper. The scores and confidence intervals are detailed first in the Figures 7–10 in table form (including ranks), and then in graphical form in Figures 11–16. In the graphs, system scores are indicated by a point, the confidence intervals by shaded areas around the point.

In all figures, we present the per-sentence normalized judgements. The normalization on a per-judge basis gave very similar ranking, only slightly less consistent with the ranking from the pairwise comparisons.

The confidence intervals are computed by bootstrap resampling for BLEU, and by standard significance testing for the manual scores, as described earlier in the paper.

Pairwise comparison is done using the sign test. Often, two systems can not be distinguished with a confidence of over 95%, so there are ranked the same. This actually happens quite frequently (more below), so that the rankings are broad estimates. For instance: if 10 systems participate, and one system does better than 3 others, worse then 2, and is not significant different from the remaining 4, its rank is in the interval 3–7.

Domain	BLEU	Fluency	Adequacy
in-domain	26.63	3.17	3.58
out-of-domain	20.37	2.74	3.08

Figure 5: Evaluation scores for in-domain and out-of-domain test sets, averaged over all systems

4.1 Close results

At first glance, we quickly recognize that many systems are scored very similar, both in terms of manual judgement and BLEU. There may be occasionally a system clearly at the top or at the bottom, but most systems are so close that it is hard to distinguish them.

In Figure 4, we displayed the number of system comparisons, for which we concluded statistical significance. For the automatic scoring method BLEU, we can distinguish three quarters of the systems. While the Bootstrap method is slightly more sensitive, it is very much in line with the sign test on text blocks.

For the manual scoring, we can distinguish only half of the systems, both in terms of fluency and adequacy. More judgements would have enabled us to make better distinctions, but it is not clear what the upper limit is. We can check, what the consequences of less manual annotation of results would have been: With half the number of manual judgements, we can distinguish about 40% of the systems, 10% less.

4.2 In-domain vs. out-of-domain

The test set included 2000 sentences from the Europarl corpus, but also 1064 sentences out-of-domain test data. Since the inclusion of out-of-domain test data was a very late decision, the participants were not informed of this. So, this was a surprise element due to practical reasons, not malice.

All systems (except for Systran, which was not tuned to Europarl) did considerably worse on out-of-domain training data. This is demonstrated by average scores over all systems, in terms of BLEU, *fluency* and *adequacy*, as displayed in Figure 5.

The manual scores are averages over the raw unnormalized scores.

Language Pair	BLEU	Fluency	Adequacy
French-English	26.09	3.25	3.61
Spanish-English	28.18	3.19	3.71
German-English	21.17	2.87	3.10
English-French	28.33	2.86	3.16
English-Spanish	27.49	2.86	3.34
English-German	14.01	3.15	3.65

Figure 6: Average scores for different language pairs. Manual scoring is done by different judges, resulting in a not very meaningful comparison.

4.3 Language pairs

It is well know that language pairs such as English-German pose more challenges to machine translation systems than language pairs such as French-English. Different sentence structure and rich target language morphology are two reasons for this.

Again, we can compute average scores for all systems for the different language pairs (Figure 6). The differences in difficulty are better reflected in the BLEU scores than in the raw un-normalized manual judgements. The easiest language pair according to BLEU (English-French: 28.33) received worse manual scores than the hardest (English-German: 14.01). This is because different judges focused on different language pairs. Hence, the different averages of manual scores for the different language pairs reflect the behaviour of the judges, not the quality of the systems on different language pairs.

4.4 Manual judgement vs. BLEU

Given the closeness of most systems and the wide over-lapping confidence intervals it is hard to make strong statements about the correlation between human judgements and automatic scoring methods such as BLEU.

We confirm the finding by Callison-Burch et al. (2006) that the rule-based system of Systran is not adequately appreciated by BLEU. In-domain Systran scores on this metric are lower than all statistical systems, even the ones that have much worse human scores. Surprisingly, this effect is much less obvious for out-of-domain test data. For instance, for out-of-domain English-French, Systran has the best BLEU and manual scores.

Our suspicion is that BLEU is very sensitive to jargon, to selecting exactly the right words, and not synonyms that human judges may appreciate as equally good. This is can not be the only explanation, since the discrepancy still holds, for instance, for out-of-domain French-English, where Systran receives among the best adequacy and fluency scores, but a worse BLEU score than all but one statistical system.

This data set of manual judgements should provide a fruitful resource for research on better automatic scoring methods.

4.5 Best systems

So, who won the competition? The best answer to this is: many research labs have very competitive systems whose performance is hard to tell apart. This is not completely surprising, since all systems use very similar technology.

For some language pairs (such as German-English) system performance is more divergent than for others (such as English-French), at least as measured by BLEU.

The statistical systems seem to still lag behind the commercial rule-based competition when translating into morphological rich languages, as demonstrated by the results for English-German and English-French.

The predominate focus of building systems that translate into English has ignored so far the difficult issues of generating rich morphology which may not be determined solely by local context.

4.6 Comments on Manual Evaluation

This is the first time that we organized a large-scale manual evaluation. While we used the standard metrics of the community, the we way presented translations and prompted for assessment differed from other evaluation campaigns. For instance, in the recent IWSLT evaluation, first fluency annotations were solicited (while withholding the source sentence), and then adequacy annotations.

Almost all annotators reported difficulties in maintaining a consistent standard for fluency and adequacy judgements, but nevertheless most did not explicitly move towards a ranking-based evaluation. Almost all annotators expressed their preference to move to a ranking-based evaluation in the future. A few pointed out that adequacy should be broken up

into two criteria: (a) are all source words covered? (b) does the translation have the same meaning, including connotations?

Annotators suggested that long sentences are almost impossible to judge. Since all long sentence translation are somewhat *muddled*, even a contrastive evaluation between systems was difficult. A few annotators suggested to break up long sentences into clauses and evaluate these separately.

Not every annotator was fluent in both the source and the target language. While it is essential to be fluent in the target language, it is not strictly necessary to know the source language, if a reference translation was given. However, ince we extracted the test corpus automatically from web sources, the reference translation was not always accurate — due to sentence alignment errors, or because translators did not adhere to a strict sentence-by-sentence translation (say, using pronouns when referring to entities mentioned in the previous sentence). Lack of correct reference translations was pointed out as a short-coming of our evaluation. One annotator suggested that this was the case for as much as 10% of our test sentences. Annotators argued for the importance of having correct and even multiple references.

It was also proposed to allow annotators to skip sentences that they are unable to judge.

5 Conclusions

We carried out an extensive manual and automatic evaluation of machine translation performance on European language pairs. While many systems had similar performance, the results offer interesting insights, especially about the relative performance of statistical and rule-based systems.

Due to many similarly performing systems, we are not able to draw strong conclusions on the question of correlation of manual and automatic evaluation metrics. The bias of automatic methods in favor of statistical systems seems to be less pronounced on out-of-domain test data.

The manual evaluation of scoring translation on a graded scale from 1–5 seems to be very hard to perform. Replacing this with an ranked evaluation seems to be more suitable. Human judges also pointed out difficulties with the evaluation of long sentences.

Acknowledgements

The manual evaluation would not have been possible without the contributions of the manual annotators: Jesus Andres Ferrer, Abhishek Arun, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Jorge Civera, Marta Ruiz Costa-jussà, Josep Maria Crego, Elsa Cubel, Chris Irwin Davis, Loic Dugast, Chris Dyer, Andreas Eisele, Cameron Fordyce, Jesús Giménez, Fabrizio Gotti, Hieu Hoang, Eric Joanis Howard Johnson, Philipp Koehn, Beata Kouchnir, Roland Kuhn, Elliott Macklovitch, Arul Menezes, Marian Olteanu, Chris Quirk, Reinhard Rapp, Fatiha Sadat, Joan Andreu Sànchez, Germán Sanchis, Michel Simard, Ashish Venugopal, and Taro Watanabe.

This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

References

Birch, A., Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Constraining the phrase-based, joint probability statistical translation model. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 154–157, New York City. Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL*.

Collins, M., Koehn, P., and Kucerova, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of ACL*.

Costa-jussà, M. R., Crego, J. M., de Gispert, A., Lambert, P., Khalilov, M., Mariño, J. B., Fonollosa, J. A. R., and Banchs, R. (2006). Talp phrase-based statistical translation system for european language pairs. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 142–145, New York City. Association for Computational Linguistics.

Crego, J. M., de Gispert, A., Lambert, P., Costa-jussà, M. R., Khalilov, M., Banchs, R., Mariño, J. B., and Fonollosa, J. A. R. (2006). N-gram-based smt system enhanced with reordering patterns. In *Proceedings on the Workshop on Statis-*

tical Machine Translation, pages 162–165, New York City. Association for Computational Linguistics.

Doddington, G. (2002). The NIST automated measure and its relation to IBM's BLEU. In *Proceedings of LREC-2002 Workshopon Machine Translation Evaluation: Human Evaluators Meet Automated Metrics*, Gran Canaria, Spain.

Eck, M. and Hori, C. (2005). Overview of the iwslt 2005 evaluation campaign. In *Proc. of the International Workshop on Spoken Language Translation*.

Giménez, J. and Màrquez, L. (2006). The ldv-combo system for smt. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 166–169, New York City. Association for Computational Linguistics.

Johnson, H., Sadat, F., Foster, G., Kuhn, R., Simard, M., Joanis, E., and Larkin, S. (2006). Portage: with smoothed phrase tables and segment choice models. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 134–137, New York City. Association for Computational Linguistics.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Lin, D. and Wu, D., editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Koehn, P. and Monz, C. (2005). Shared task: Statistical machine translation between European languages. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 119–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Li, A. (2005). Results of the 2005 NIST machine translation evaluation. In *Machine Translation Workshop*.

Menezes, A., Toutanova, K., and Quirk, C. (2006). Microsoft research treelet translation system: Naacl 2006 europarl evaluation. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 158–161, New York City. Association for Computational Linguistics.

Olteanu, M., Davis, C., Volosen, I., and Moldovan, D. (2006a). Phramer - an open source statistical phrase-based translator. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 146–149, New York City. Association for Computational Linguistics.

Olteanu, M., Suriyentrakorn, P., and Moldovan, D. (2006b). Language models and reranking for machine translation. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 150–153, New York City. Association for Computational Linguistics.

Patry, A., Gotti, F., and Langlais, P. (2006). Mood at work: Ramses versus pharaoh. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 126–129, New York City. Association for Computational Linguistics.

Przybocki, M. (2004). NIST machine translation 2004 evaluation – summary of results. In *Machine Translation Evaluation Workshop*.

Riezler, S. and Maxwell, J. T. (2005). On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

Sánchez, J. A. and Benedí, J. M. (2006). Stochastic inversion transduction grammars for obtaining word phrases for phrase-based statistical machine translation. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 130–133, New York City. Association for Computational Linguistics.

Watanabe, T., Tsukada, H., and Isozaki, H. (2006). Ntt system description for the wmt2006 shared task. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 122–125, New York City. Association for Computational Linguistics.

Zollmann, A. and Venugopal, A. (2006). Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, New York City. Association for Computational Linguistics.

French-English (In Domain)

	Adequacy (rank)	Fluency (rank)	BLEU (rank)
upc-jmc	+0.19±0.08 (1-7)	+0.09±0.08 (1-8)	30.42±0.86 (1-6)
lcc	+0.14±0.07 (1-6)	+0.13±0.06 (1-7)	30.81±0.85 (1-4)
utd	+0.13±0.08 (1-7)	+0.14±0.07 (1-6)	30.53±0.87 (2-7)
upc-mr	+0.13±0.08 (1-8)	+0.13±0.07 (1-6)	30.33±0.88 (1-7)
nrc	+0.12±0.10 (1-7)	+0.06±0.11 (2-6)	29.62±0.84 (8)
ntt	+0.11±0.08 (1-8)	+0.14±0.08 (2-8)	30.72±0.87 (1-7)
cmu	+0.10±0.08 (3-7)	+0.05±0.07 (4-8)	30.18±0.80 (2-7)
rali	-0.02±0.08 (5-8)	+0.00±0.08 (3-9)	30.39±0.91 (3-7)
systran	-0.08±0.09 (9)	-0.17±0.09 (8-9)	21.44±0.65 (10)
upv	-0.76±0.09 (10)	-0.52±0.09 (10)	24.10±0.89 (9)

Spanish-English (In Domain)

	Adequacy (rank)	Fluency (rank)	BLEU (rank)
upc-jmc	+0.15±0.08 (1-7)	+0.18±0.08 (1-6)	31.01±0.97 (1-5)
ntt	+0.10±0.08 (1-7)	+0.10±0.08 (1-8)	31.29±0.88 (1-5)
lcc	+0.08±0.07 (1-8)	+0.04±0.06 (2-8)	31.46±0.87 (1-4)
utd	+0.08±0.06 (1-8)	+0.08±0.07 (2-7)	31.10±0.89 (1-5)
nrc	+0.06±0.10 (2-8)	+0.08±0.07 (1-9)	30.04±0.79 (6)
upc-mr	+0.06±0.07 (1-8)	+0.08±0.07 (1-6)	29.43±0.83 (7)
uedin-birch	+0.03±0.11 (1-8)	-0.07±0.15 (2-10)	29.01±0.81 (8)
rali	+0.00±0.07 (3-9)	-0.02±0.07 (3-9)	30.80±0.87 (2-5)
upc-jg	-0.10±0.07 (7-9)	-0.11±0.07 (6-9)	28.03±0.83 (9)
upv	-0.45±0.10 (10)	-0.41±0.10 (9-10)	23.91±0.83 (10)

German-English (In Domain)

	Adequacy (rank)	Fluency (rank)	BLEU (rank)
uedin-phi	+0.30±0.09 (1-2)	+0.33±0.08 (1)	27.30±0.86 (1)
lcc	+0.15±0.07 (2-7)	+0.12±0.07 (2-7)	25.97±0.81 (2)
nrc	+0.12±0.07 (2-7)	+0.14±0.07 (2-6)	24.54±0.80 (5-7)
utd	+0.08±0.07 (3-7)	+0.01±0.08 (2-8)	25.44±0.85 (3-4)
ntt	+0.07±0.08 (2-9)	+0.06±0.09 (2-8)	25.64±0.83 (3-4)
upc-mr	+0.00±0.09 (3-9)	-0.21±0.09 (6-9)	23.68±0.79 (8)
rali	-0.01±0.06 (4-9)	+0.00±0.07 (3-9)	24.60±0.80 (5-7)
upc-jmc	-0.02±0.09 (2-9)	-0.04±0.09 (3-9)	24.43±0.86 (5-7)
systran	-0.05±0.10 (3-9)	-0.05±0.09 (3-9)	15.86±0.59 (10)
upv	-0.55±0.09 (10)	-0.38±0.08 (10)	18.08±0.77 (9)

Figure 7: Evaluation of translation to English on in-domain test data

English-French (In Domain)

	Adequacy (rank)	Fluency (rank)	BLEU (rank)
nrc	+0.08±0.09 (1-5)	+0.09±0.09 (1-5)	31.75±0.83 (1-6)
upc-mr	+0.08±0.08 (1-4)	+0.04±0.07 (1-5)	31.50±0.76 (1-6)
upc-jmc	+0.03±0.09 (1-6)	+0.02±0.08 (1-6)	31.75±0.78 (1-5)
systran	-0.01±0.12 (2-7)	+0.06±0.12 (1-6)	25.07±0.71 (7)
utd	-0.03±0.07 (3-7)	-0.05±0.07 (3-7)	31.42±0.85 (3-6)
rali	-0.08±0.09 (1-7)	-0.09±0.09 (2-7)	31.79±0.85 (1-6)
ntt	-0.09±0.09 (4-7)	-0.06±0.08 (4-7)	31.92±0.84 (1-5)

English-Spanish (In Domain)

	Adequacy (rank)	Fluency (rank)	BLEU (rank)
ms	+0.23±0.09 (1-5)	+0.13±0.09 (1-7)	29.76±0.82 (7-8)
upc-mr	+0.20±0.09 (1-4)	+0.17±0.09 (1-5)	31.06±0.86 (1-4)
utd	+0.18±0.08 (1-5)	+0.15±0.08 (1-6)	30.73±0.90 (1-4)
nrc	+0.12±0.09 (2-7)	+0.17±0.08 (1-6)	29.97±0.86 (5-6)
ntt	+0.10±0.09 (3-7)	+0.14±0.08 (1-6)	30.93±0.85 (1-4)
upc-jmc	+0.04±0.10 (2-7)	+0.01±0.08 (2-7)	30.44±0.86 (1-4)
rali	-0.05±0.08 (5-8)	-0.03±0.08 (6-8)	29.38±0.85 (5-6)
uedin-birch	-0.18±0.14 (6-9)	-0.17±0.13 (6-10)	28.49±0.87 (7-8)
upc-jg	-0.32±0.11 (9)	-0.37±0.09 (8-10)	27.46±0.78 (9)
upv	-0.83±0.15 (9-10)	-0.59±0.15 (8-10)	23.17±0.73 (10)

English-German (In Domain)

	Adequacy (rank)	Fluency (rank)	BLEU (rank)
upc-mr	+0.28±0.08 (1-3)	+0.14±0.08 (1-5)	17.24±0.81 (3-5)
ntt	+0.19±0.08 (1-5)	+0.09±0.06 (2-6)	18.15±0.89 (1-3)
upc-jmc	+0.17±0.08 (1-5)	+0.13±0.08 (1-4)	17.73±0.81 (1-3)
nrc	+0.17±0.08 (2-4)	+0.11±0.08 (1-5)	17.52±0.78 (4-5)
rali	+0.08±0.10 (3-6)	+0.03±0.09 (2-6)	17.93±0.85 (1-4)
systran	-0.08±0.11 (5-6)	+0.00±0.10 (3-6)	9.84±0.52 (7)
upv	-0.84±0.12 (7)	-0.51±0.10 (7)	13.37±0.78 (6)

Figure 8: Evaluation of translation from English on in-domain test data

French-English (Out of Domain)

	Adequacy (rank)	Fluency (rank)	BLEU (rank)
upc-jmc	+0.23±0.09 (1-5)	+0.13±0.11 (1-8)	21.79±0.92 (1-4)
cmu	+0.22±0.11 (1-8)	+0.13±0.09 (1-9)	21.15±0.86 (4-7)
systran	+0.19±0.15 (1-8)	+0.15±0.14 (1-7)	19.42±0.82 (9)
lcc	+0.13±0.12 (1-9)	+0.11±0.11 (1-9)	21.77±0.88 (1-5)
upc-mr	+0.12±0.12 (2-8)	+0.11±0.10 (1-7)	21.95±0.94 (1-3)
utd	+0.04±0.10 (1-9)	+0.01±0.10 (1-8)	21.39±0.94 (3-7)
ntt	-0.02±0.12 (3-9)	+0.08±0.11 (1-9)	21.34±0.85 (3-7)
nrc	-0.03±0.14 (3-8)	+0.00±0.11 (3-9)	21.15±0.86 (3-7)
rali	-0.09±0.12 (4-9)	-0.10±0.11 (5-9)	20.17±0.85 (8)
upv	-0.76±0.16 (10)	-0.58±0.14 (10)	15.55±0.79 (10)

Spanish-English (Out of Domain)

	Adequacy (rank)	Fluency (rank)	BLEU (rank)
upc-jmc	+0.28±0.10 (1-2)	+0.17±0.10 (1-6)	27.92±0.94 (1-3)
uedin-birch	+0.25±0.16 (1-7)	+0.18±0.19 (1-6)	25.20±0.91 (5-8)
nrc	+0.18±0.16 (2-8)	+0.09±0.09 (1-8)	25.40±0.94 (5-7)
ntt	+0.11±0.10 (2-7)	+0.17±0.10 (2-6)	26.85±0.89 (3-4)
upc-mr	+0.08±0.11 (2-8)	+0.10±0.10 (1-7)	25.62±0.87 (5-8)
lcc	+0.04±0.10 (4-9)	+0.07±0.11 (3-7)	27.18±0.92 (1-4)
utd	+0.03±0.11 (2-9)	+0.03±0.10 (2-8)	27.41±0.96 (1-3)
upc-jg	-0.09±0.11 (4-9)	-0.09±0.09 (7-9)	23.42±0.87 (9)
rali	-0.09±0.11 (4-9)	-0.15±0.11 (6-9)	25.03±0.91 (6-8)
upv	-0.63±0.14 (10)	-0.47±0.11 (10)	19.17±0.78 (10)

German-English (Out of Domain)

	Adequacy (rank)	Fluency (rank)	BLEU (rank)
systran	+0.30±0.12 (1-4)	+0.21±0.12 (1-4)	15.56±0.71 (7-9)
uedin-phi	+0.22±0.09 (1-6)	+0.21±0.10 (1-7)	18.87±0.84 (1)
lcc	+0.18±0.10 (1-6)	+0.20±0.10 (1-7)	17.96±0.79 (2-3)
utd	+0.08±0.09 (2-7)	+0.07±0.08 (2-6)	16.97±0.76 (4-6)
ntt	+0.07±0.12 (1-9)	+0.21±0.13 (1-7)	17.37±0.76 (3-5)
nrc	+0.04±0.10 (3-8)	+0.04±0.09 (2-8)	15.93±0.76 (7-8)
upc-mr	+0.02±0.10 (4-8)	-0.11±0.09 (6-8)	16.89±0.79 (4-6)
upc-jmc	-0.01±0.10 (4-8)	-0.04±0.11 (3-9)	17.57±0.80 (2-5)
rali	-0.14±0.08 (8-9)	-0.14±0.08 (8-9)	15.22±0.69 (8-9)
upv	-0.64±0.11 (10)	-0.54±0.09 (10)	11.78±0.71 (10)

Figure 9: Evaluation of translation to English on out-of-domain test data

English-French (Out of Domain)

	Adequacy (rank)	Fluency (rank)	BLEU (rank)
systran	+0.50±0.20 (1)	+0.41±0.18 (1)	25.31±0.88 (1)
upc-jmc	+0.09±0.11 (2-5)	+0.09±0.11 (2-4)	23.30±0.75 (2-6)
upc-mr	+0.09±0.11 (2-4)	+0.04±0.09 (2-4)	23.21±0.75 (2-6)
utd	-0.02±0.11 (2-6)	-0.05±0.09 (2-6)	22.79±0.86 (7)
rali	-0.12±0.12 (4-7)	-0.17±0.12 (5-7)	23.34±0.89 (2-6)
nrc	-0.13±0.13 (4-7)	-0.16±0.10 (4-7)	23.66±0.91 (2-5)
ntt	-0.23±0.12 (4-7)	-0.06±0.10 (4-7)	22.99±0.96 (3-6)

English-Spanish (Out of Domain)

	Adequacy (rank)	Fluency (rank)	BLEU (rank)
upc-mr	+0.35±0.11 (1-3)	+0.19±0.10 (1-6)	26.62±0.92 (1-2)
ms	+0.33±0.16 (1-7)	+0.15±0.13 (1-8)	26.15±0.88 (6-7)
utd	+0.21±0.13 (2-6)	+0.13±0.11 (1-7)	25.26±0.78 (3-5)
nrc	+0.18±0.12 (1-6)	+0.07±0.11 (2-7)	25.58±0.85 (3-5)
upc-jmc	+0.17±0.15 (2-7)	+0.24±0.12 (1-6)	25.59±0.95 (3-5)
ntt	+0.12±0.13 (2-7)	+0.12±0.13 (1-7)	26.52±0.90 (1-2)
rali	-0.17±0.16 (6-8)	-0.05±0.13 (4-8)	24.03±0.83 (6-8)
uedin-birch	-0.36±0.24 (6-10)	-0.16±0.16 (5-9)	23.18±0.88 (7-8)
upc-jg	-0.45±0.13 (8-9)	-0.42±0.10 (9-10)	22.04±0.84 (9)
upv	-1.09±0.21 (9)	-0.64±0.19 (8-9)	16.83±0.72 (10)

English-German (Out of Domain)

	Adequacy (rank)	Fluency (rank)	BLEU (rank)
systran	+0.47±0.15 (1)	+0.39±0.15 (1-2)	10.78±0.69 (1-6)
upc-mr	+0.31±0.13 (2-3)	+0.21±0.11 (1-3)	10.96±0.70 (1-5)
upc-jmc	+0.22±0.14 (2-3)	+0.01±0.10 (3-6)	10.64±0.66 (1-6)
rali	+0.13±0.12 (4-6)	-0.06±0.10 (4-6)	10.57±0.65 (1-6)
nrc	+0.00±0.11 (4-6)	+0.05±0.09 (2-6)	10.64±0.65 (2-6)
ntt	-0.03±0.12 (4-6)	+0.08±0.11 (3-5)	10.51±0.64 (1-6)
upv	-0.94±0.13 (7)	-0.57±0.10 (7)	6.55±0.53 (7)

Figure 10: Evaluation of translation from English on out-of-domain test data

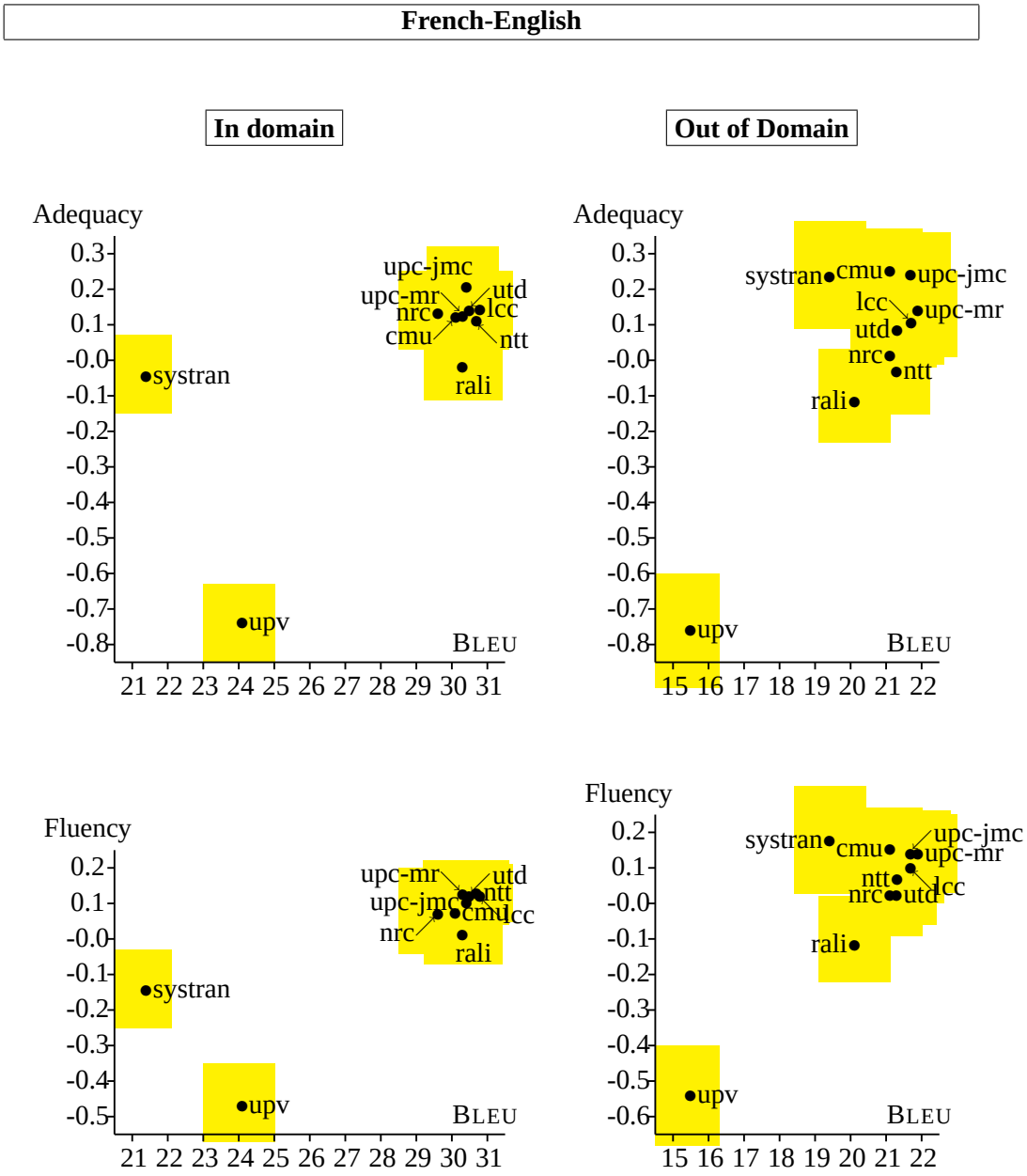


Figure 11: Correlation between manual and automatic scores for French-English

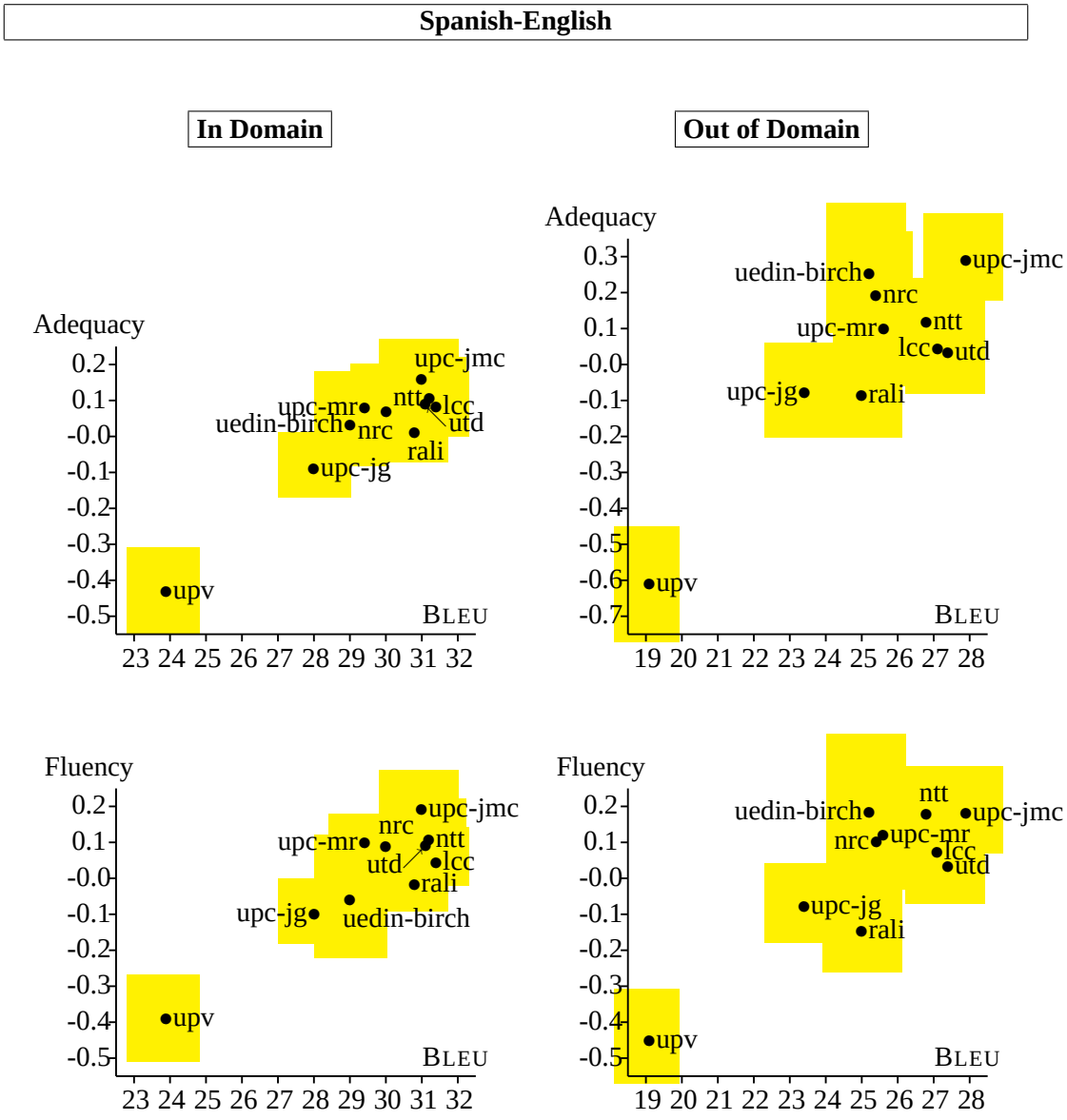


Figure 12: Correlation between manual and automatic scores for Spanish-English

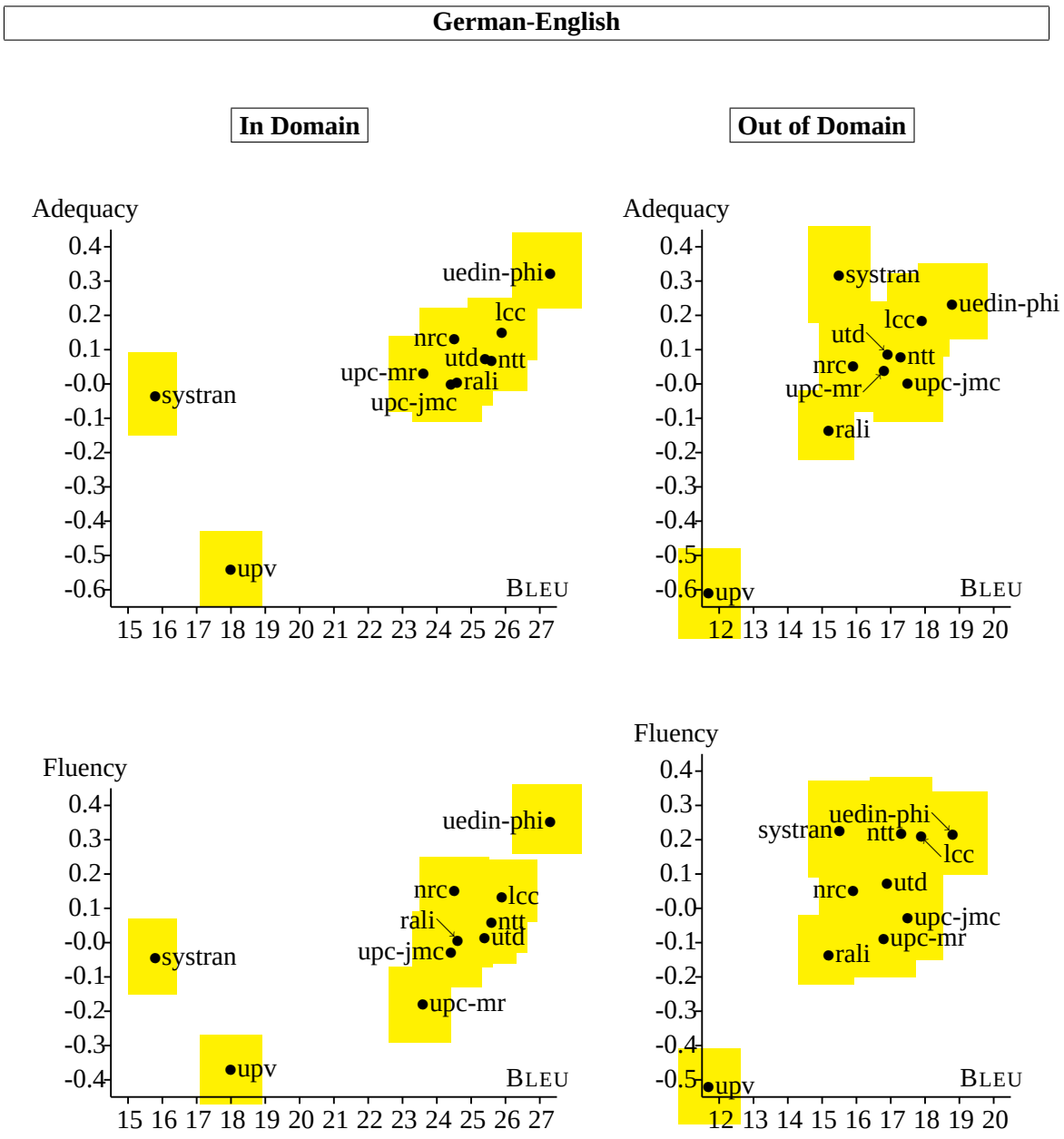


Figure 13: Correlation between manual and automatic scores for German-English

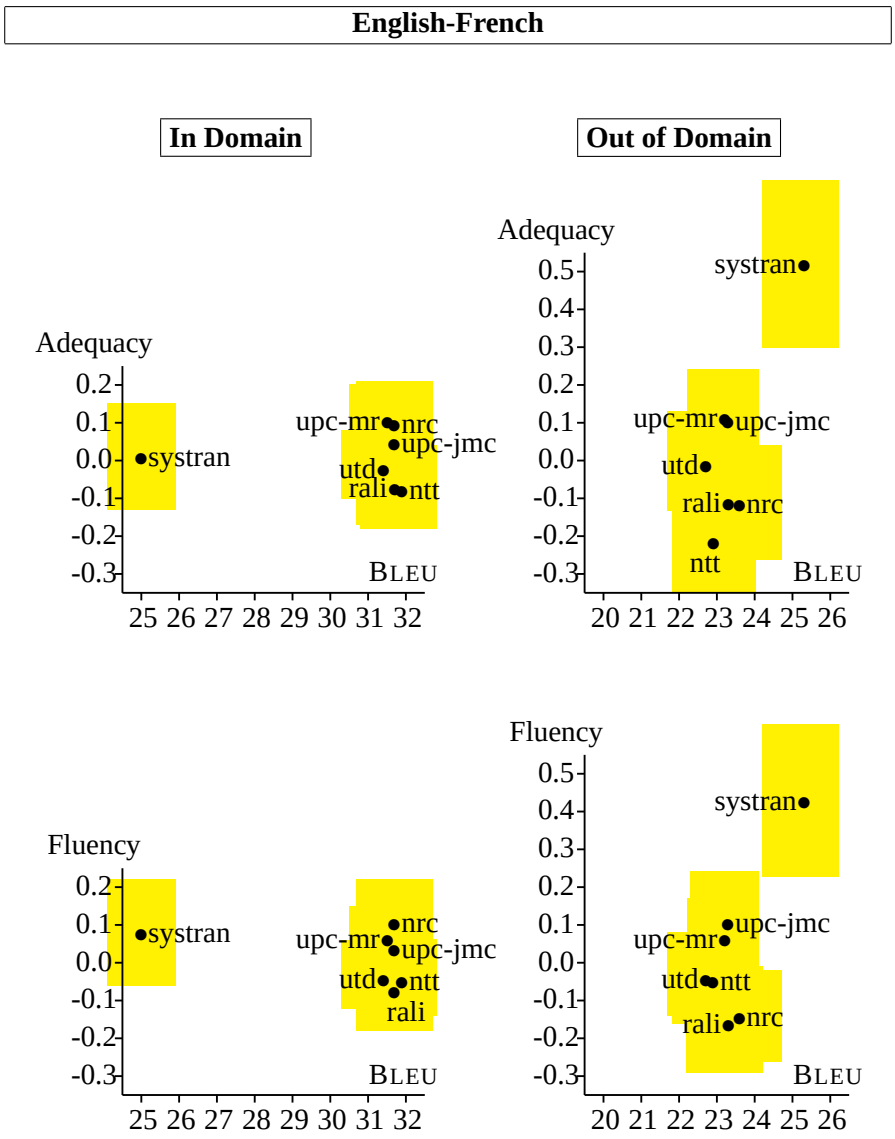


Figure 14: Correlation between manual and automatic scores for English-French

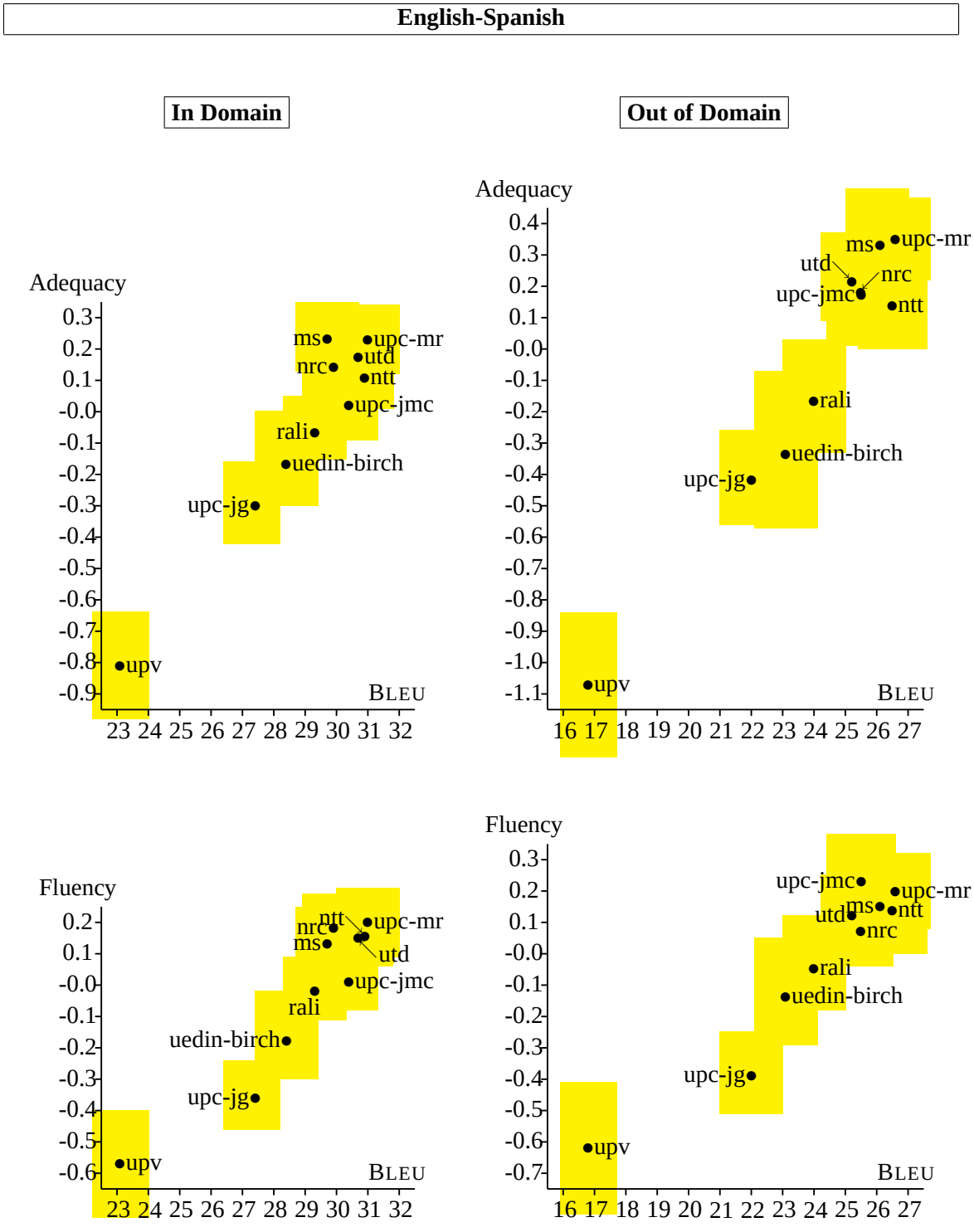


Figure 15: Correlation between manual and automatic scores for English-Spanish

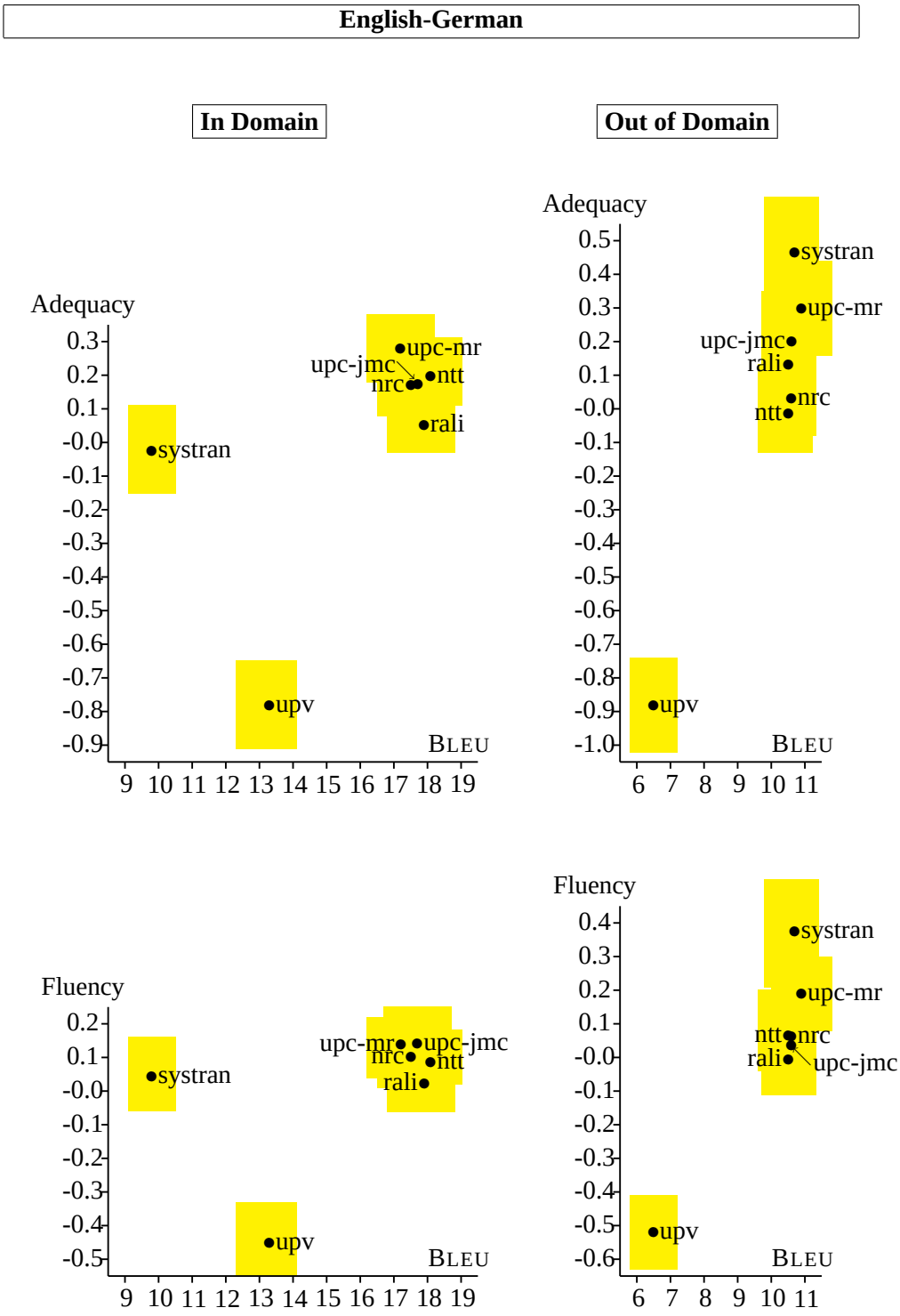


Figure 16: Correlation between manual and automatic scores for English-German